香港中文大學
The Chinese University of Hong Kong

香港中文大學醫學院
Faculty of Medicine
The Chinese University of Hong Kong

The Language of Life: Natural Language Processing in Microbiology

Joint Graduate Seminar
Supervisor: Prof. Margaret Ip
Co-Supervisors: Prof. Sishuo Wang, Prof. Guoping Zhao, Dr. Mingjing Luo
Year 2 PhD Student: Zheng Ruiqi
Date: December 10, 2024
Department: Microbiology

CONTENTS

01 Introduction to Natural Language Processing (NLP)

- What is Natural Language Processing (NLP)?
- Basic Task of NLP
- How Does NLP Work?
- An evolution process of the five generations of language models (LM)

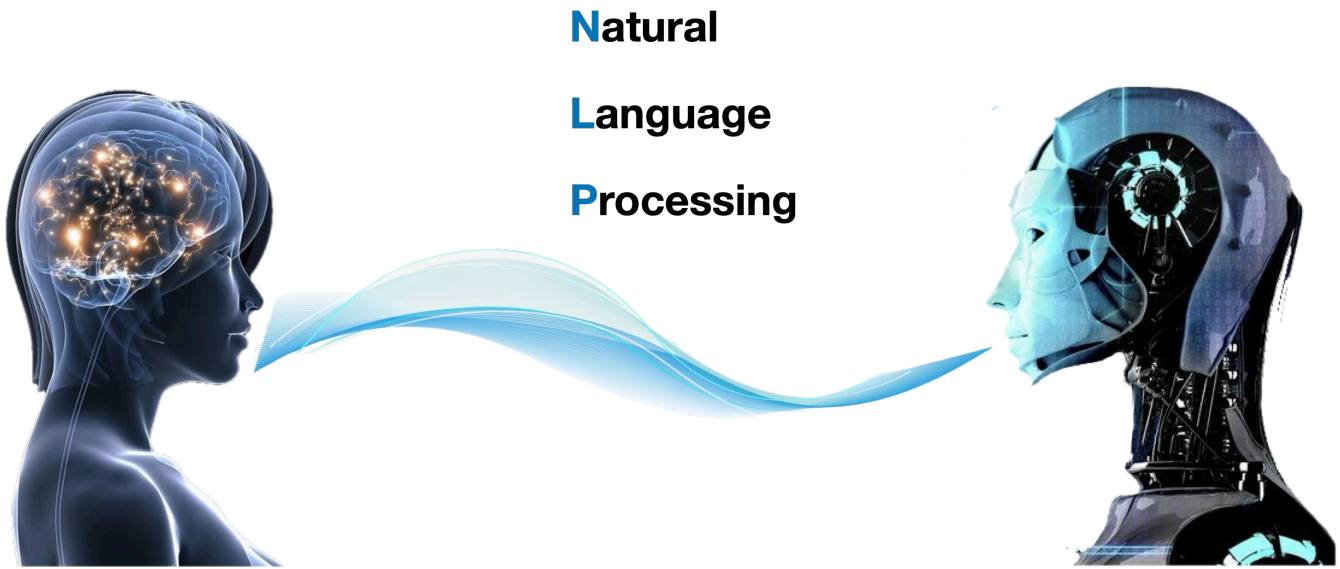
02 NLP in Microbiology

- Protein Language Modeling
- NLP transformer architecture
- APPLICATIONS
- DNA/Genomic Language Modeling
- APPLICATIONS

03 Future Directions

04 Takeaways

What is Natural Language Processing (NLP)?



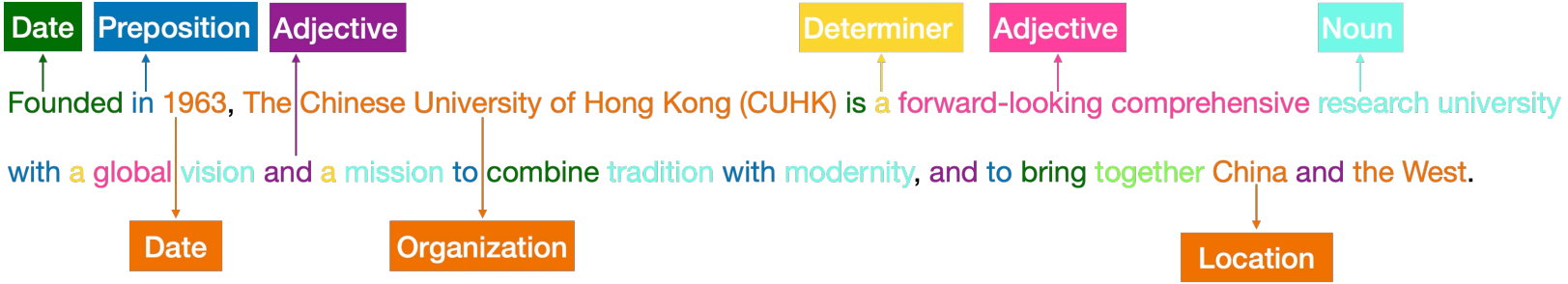
NLP is the process that helps Artificial Intelligence (AI) understand language rules and grammar. By programming, we enable AI to create complex models that represent these rules and use them to complete specific tasks.

Basic Task of NLP

Information Extraction

- Speech Tagging

Identifying parts of speech in sentences, such as nouns, verbs, adjectives, and more.



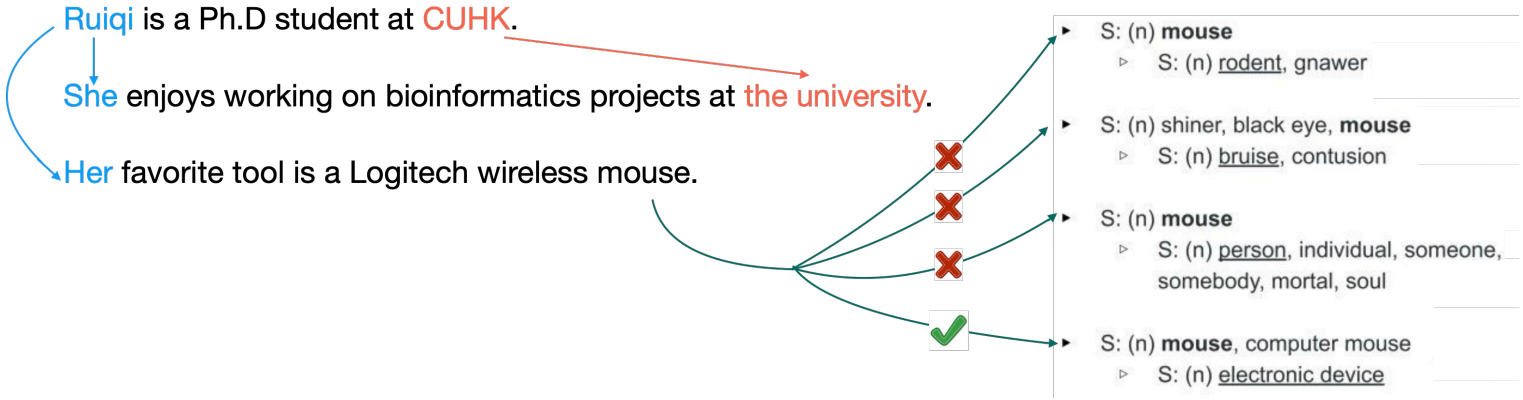
- Named Entity Recognition

Identifying multiple instances of nouns, such as names, organizations, dates, and locations.

Understanding Tasks

- Word Sense Disambiguation

Determining the exact meaning of a word with multiple meanings in a specific context.

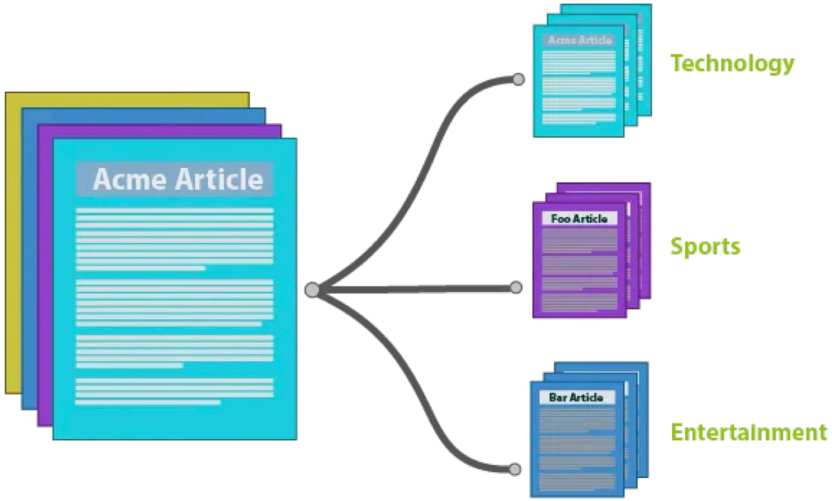


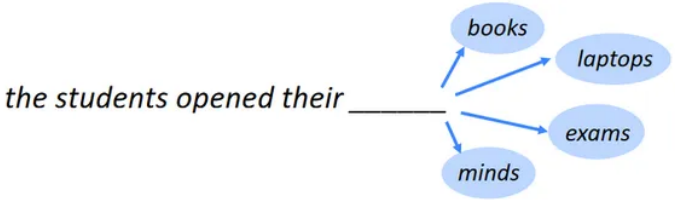
- Co-reference Resolution

Finding all expressions that refer to the same entity within a text.

Classification and Prediction

Classification Task: Assigning text to predefined categories.





Prediction Task: Predicting a specific outcome based on text data.

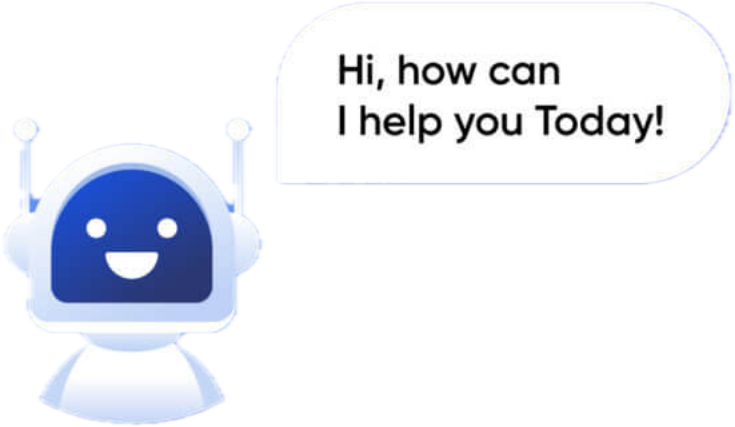
Machine Learning, NLP: Text Classification using scikit-learn, python and NLTK.

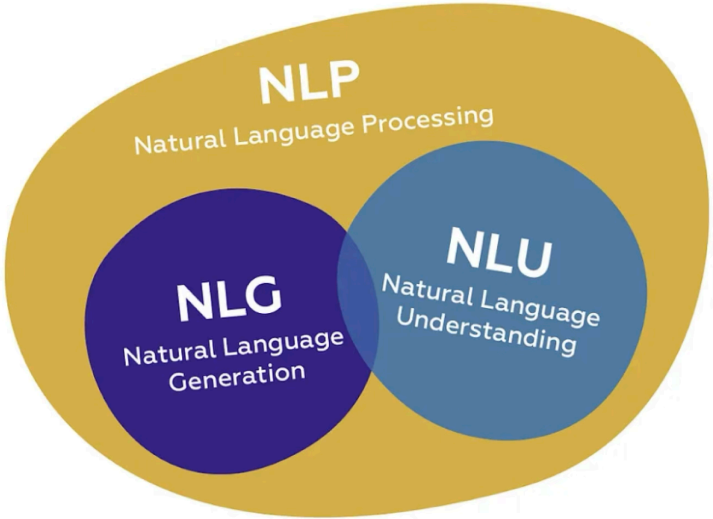
Generation Tasks

Generation Tasks and Understanding Tasks are core tasks in NLP.

NLU: Using grammar and meaning analysis to understand the meaning of text or speech.

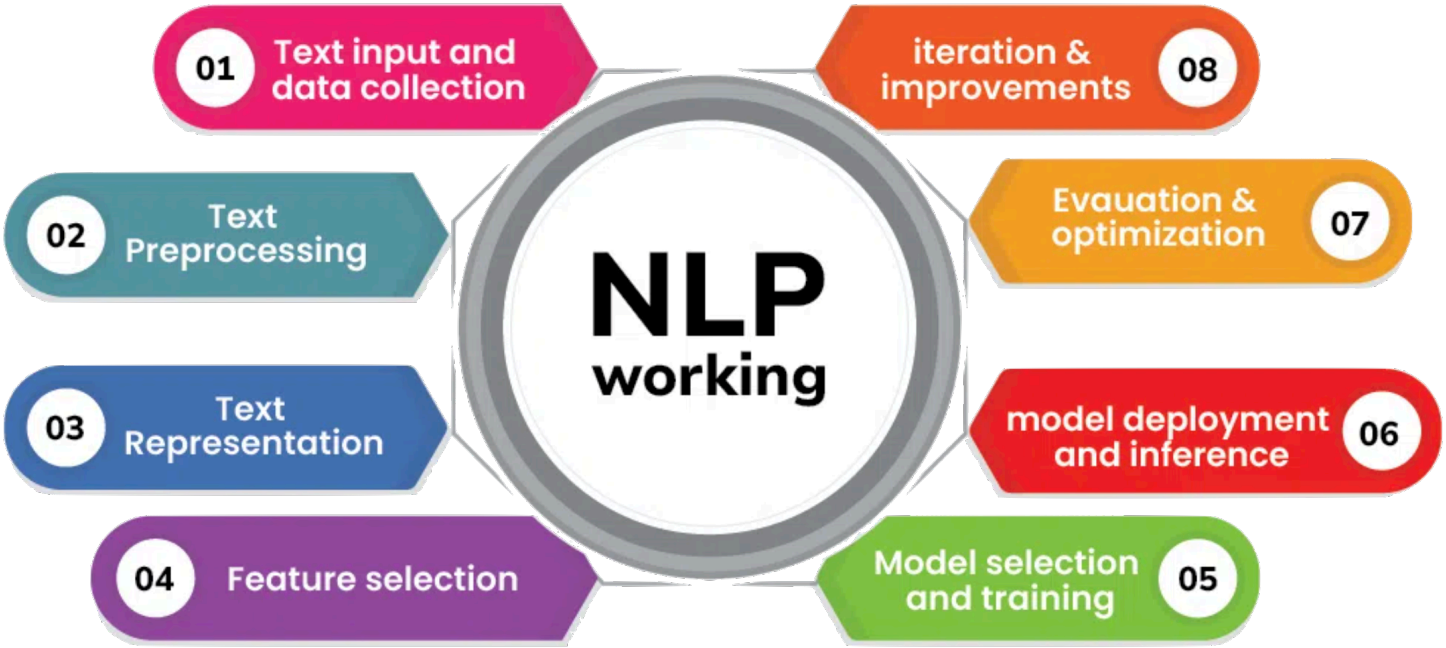
NLG: Generating text and speech from data input.



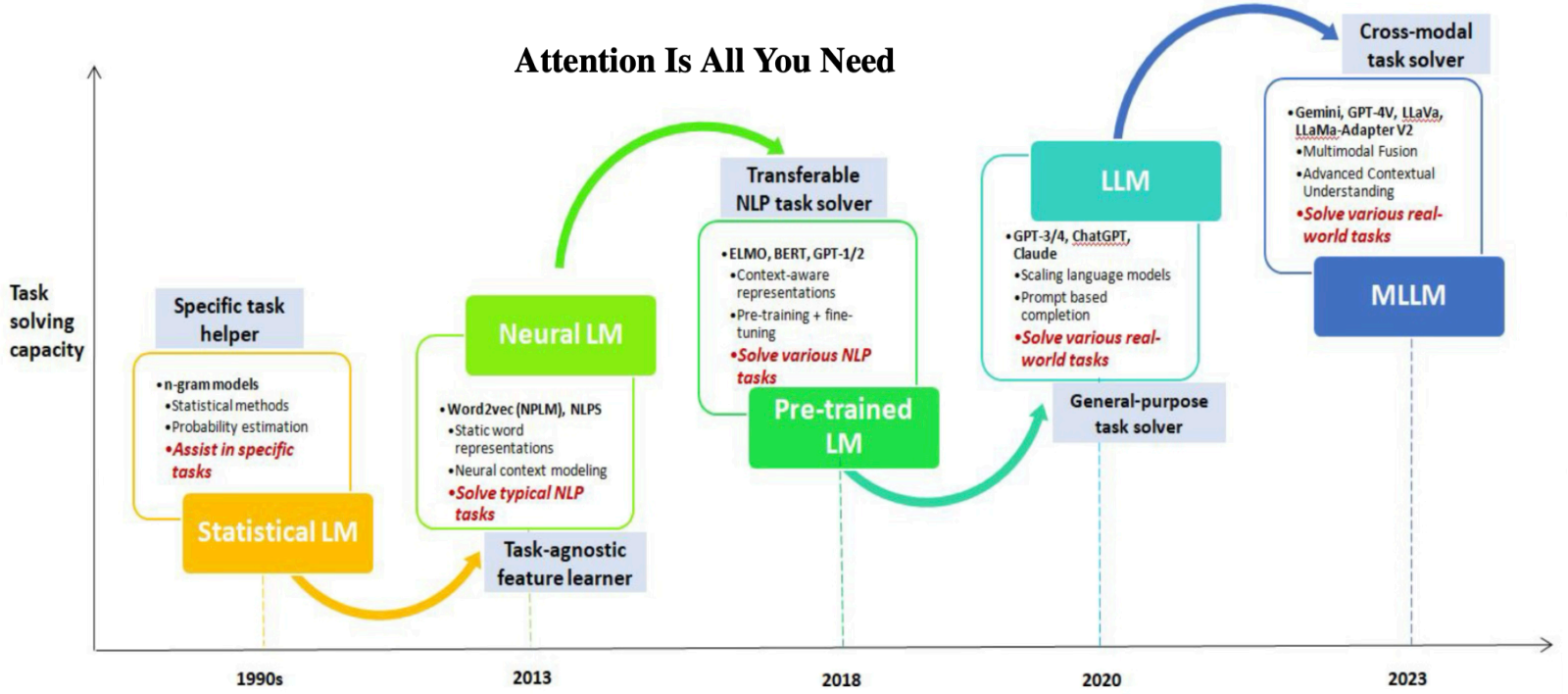


Natural Language Generation-Genspark

How Does NLP Work?

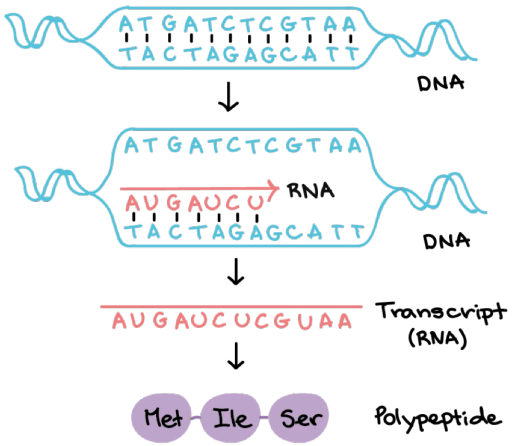


An evolution process of the five generations of language models (LM)



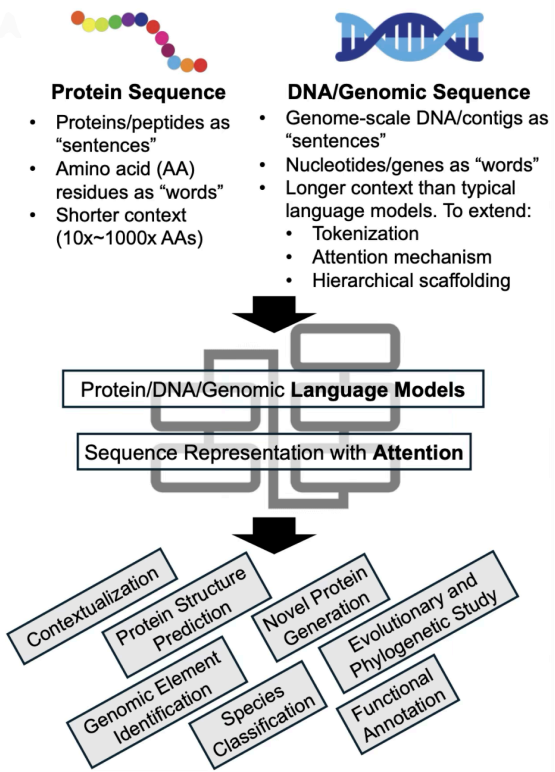
Red Teaming for Multimodal Large Language Models: A Survey

NLP in Microbiology



Microbial genomic elements are organized as:

- sequences of nucleotide base pairs (for genomic DNA)
- amino acids (AA, for proteins).



The complex dependency structure of protein/gene-level or genomic-scale sequences can be modeled by language model techniques

DNA CODE OF LIFE

Recent advances in deep learning and language models for studying the microbiome

Protein Language Modeling

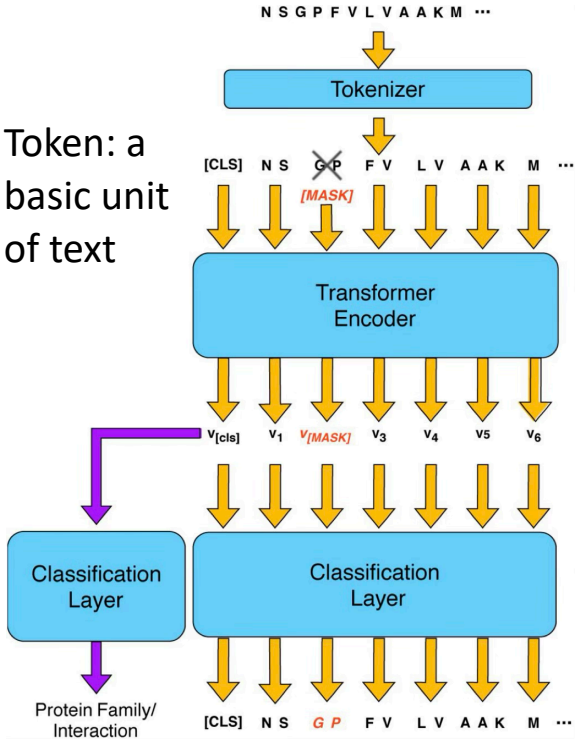
Protein language models fit well within transformer context lengths, as microbial proteins usually contain fewer than 1,000 amino acids (tokens).
Protein language models are used for designing and predicting individual proteins.

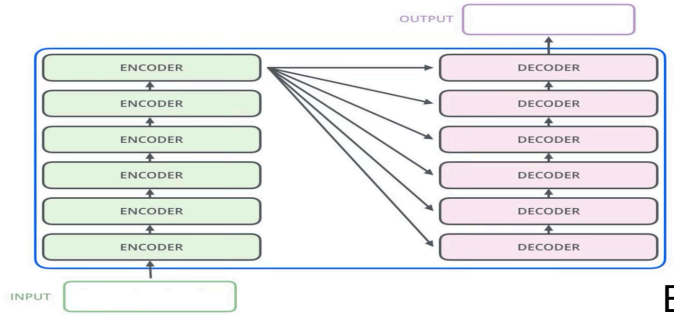
Tasks:
Identifying specific functional regions or domains (Information Extraction)
Novel protein generation (Generation Tasks)
Function and structure prediction (Classification and Prediction)

Datasets:
UniProt
Gene Ontology (GO)
Protein Data Bank (PDB)
PeptideAtlas

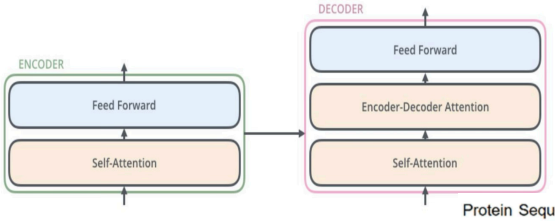
NLP transformer architecture

Token: a basic unit of text



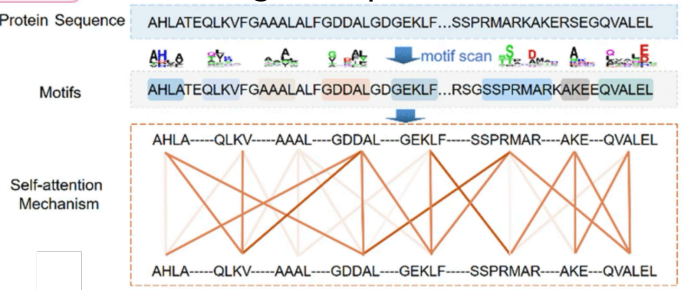


Transformer architecture



Encoder: Produce an representation for each token in the input sequence.
Decoder:Produce an representation for each word in the target sequence.

Attention mechanism: understand the relationships between words.

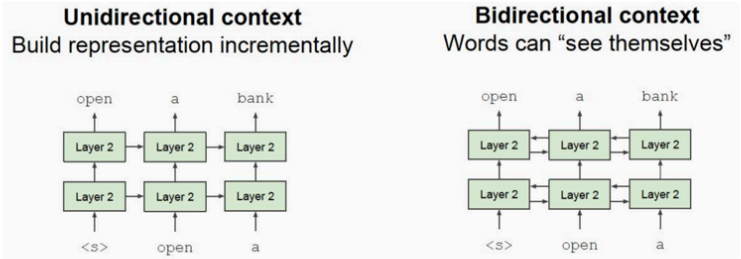


SelfAT-Fold: Protein Fold Recognition Based on Residue-Based and Motif-Based Self-Attention Networks

ProteinBERT: a universal deep-learning model of protein sequence and function

Pre-trained Language Models(PLMs): language models having powerful transferability for other NLP tasks.

BERT = Bidirectional Encoder Representations from Transformers



Lily went to the store to___

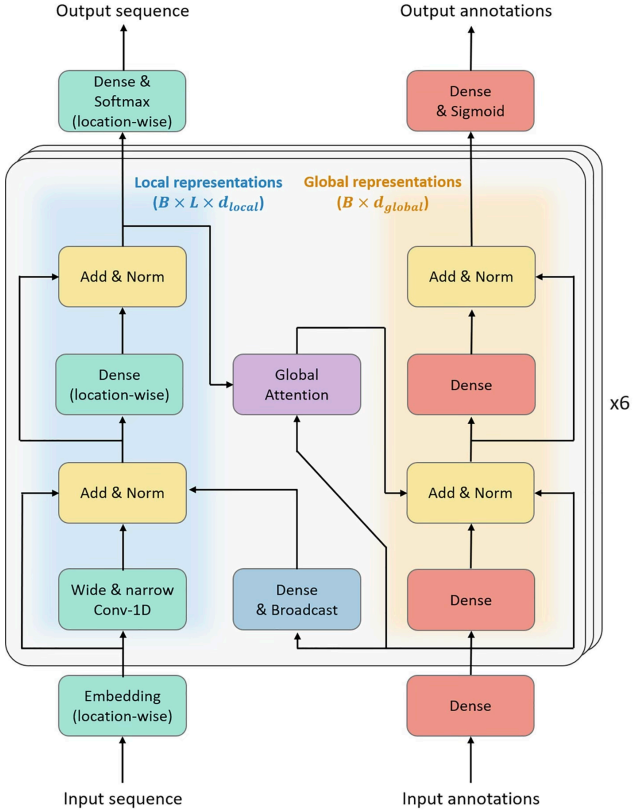
Lily went to the store to buy___ a pencil.

With pre-training, bigger == better,
without clear limits.

ProteinBERT: a universal deep-learning model of protein sequence and function

ProteinBERT architecture：

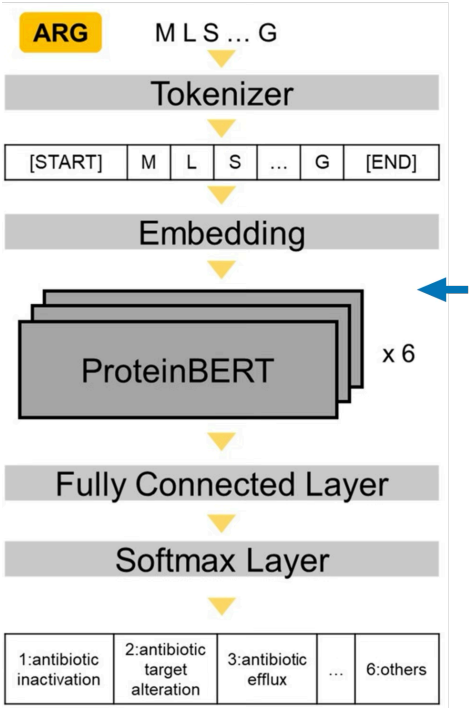
- Supports both local (sequential) and global data, unlike standard Transformers.
- Comprises six transformer-like blocks for manipulating local and global representations.



ProteinBERT provides rapid inference and effective training with limited labeled data (annotation), while also maintaining a smaller model size

Nadav Brandes.ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics, Volume 38, Issue 8, March 2022.

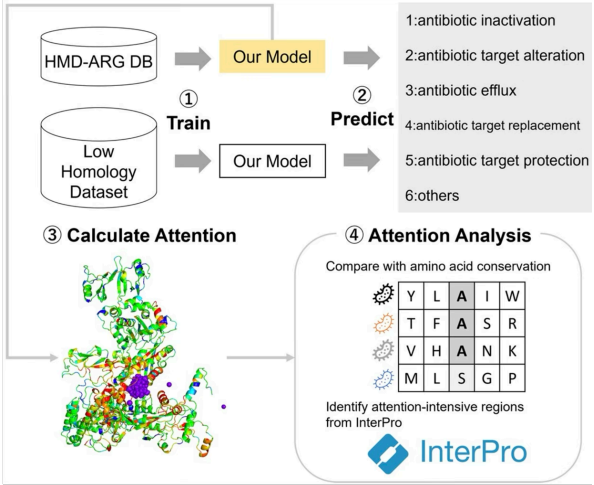
Prediction of antibiotic resistance mechanisms using ProteinBERT



Existing methods struggle to accurately predict resistance mechanisms for ARGs with low similarity to known sequences and lack sufficient interpretability of the prediction models.

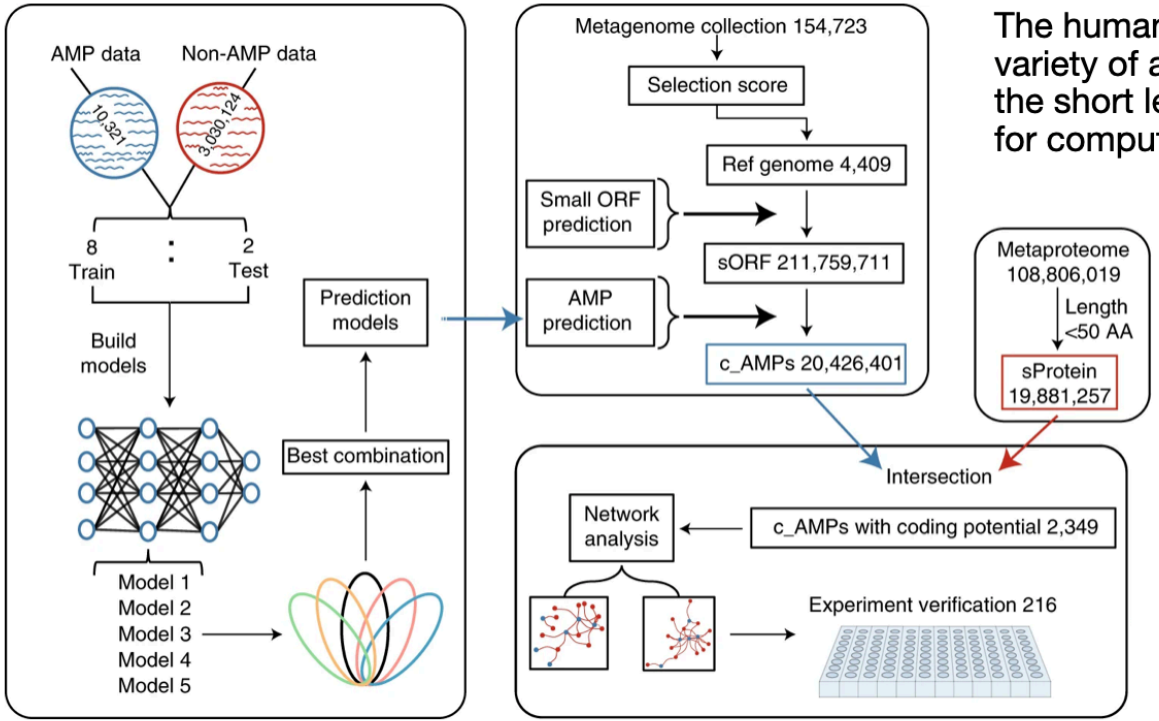
This model is based on ProteinBERT and features an input layer for ARG sequences and an output layer that predicts six resistance mechanism labels

The model was fine-tuned using the HMD-ARG DB and a custom low-homology dataset, followed by attention analysis



Kanami Yagimoto, Shion Hosoda, Miwa Sato, Michiaki Hamada, Prediction of antibiotic resistance mechanisms using a protein language model, Bioinformatics, Volume 40, Issue 10, October 2024

Identification of antimicrobial peptides from the human gut microbiome



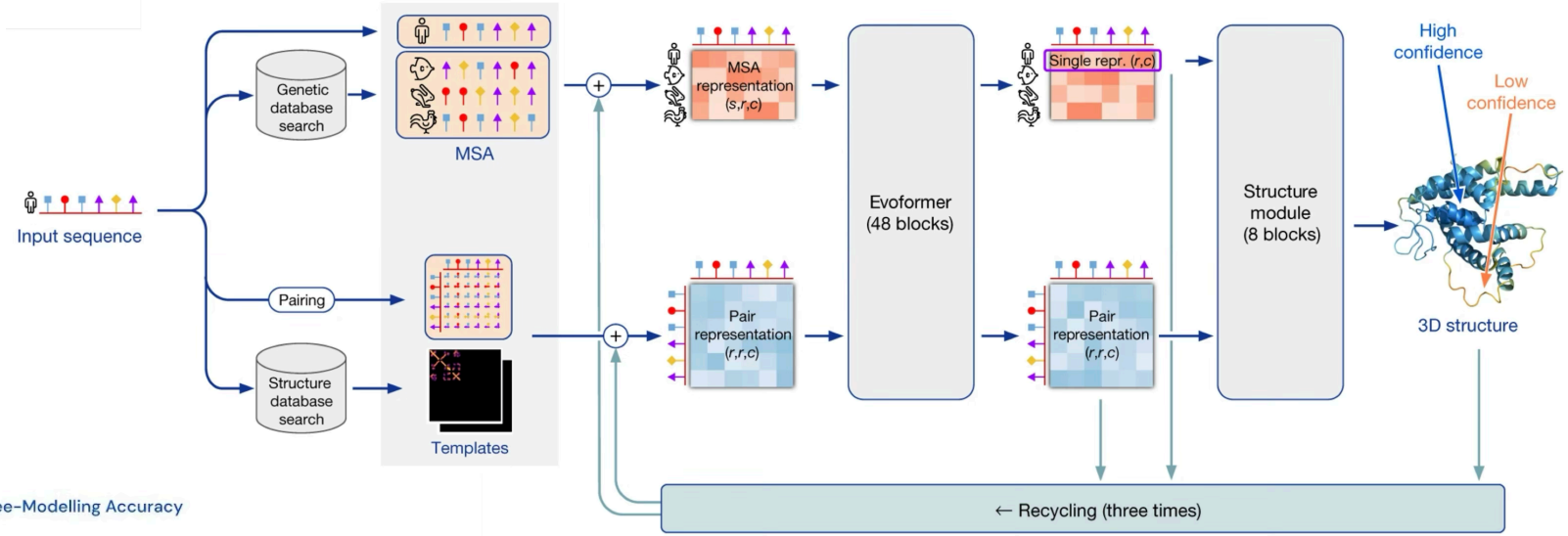
The human gut microbiome encodes a large variety of antimicrobial peptides (AMPs), but the short lengths of AMPs pose a challenge for computational prediction.

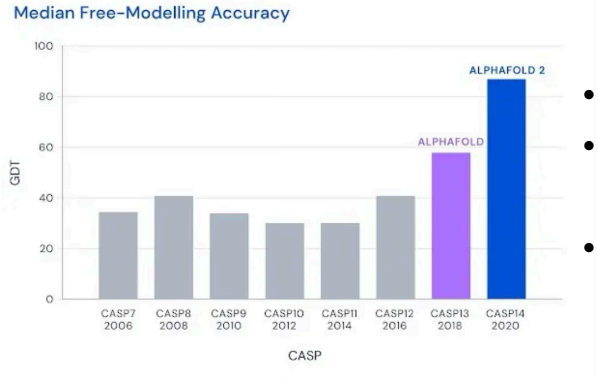
Combined multiple natural language processing neural network models: LSTM, Attention and BERT

This model effectively identifies these short peptides by treating peptide sequences as text and utilizing deep learning techniques, thereby overcoming the limitations of traditional methods.

Ma, Y., Guo, Z., Xia, B. et al. Identification of antimicrobial peptides from the human gut microbiome using deep learning. Nat Biotechnol 40, 921–931 (2022).

AlphaFold：One of the most inspiring research results!





- Predict 3D structure with the help of molecular dynamics
- MSA + EvoFormer +End2end training: perfect combination for biomedical knowledge and NLP technique
- A breakthrough for the 3D structure prediction accuracy (comparable to human level)

Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021)

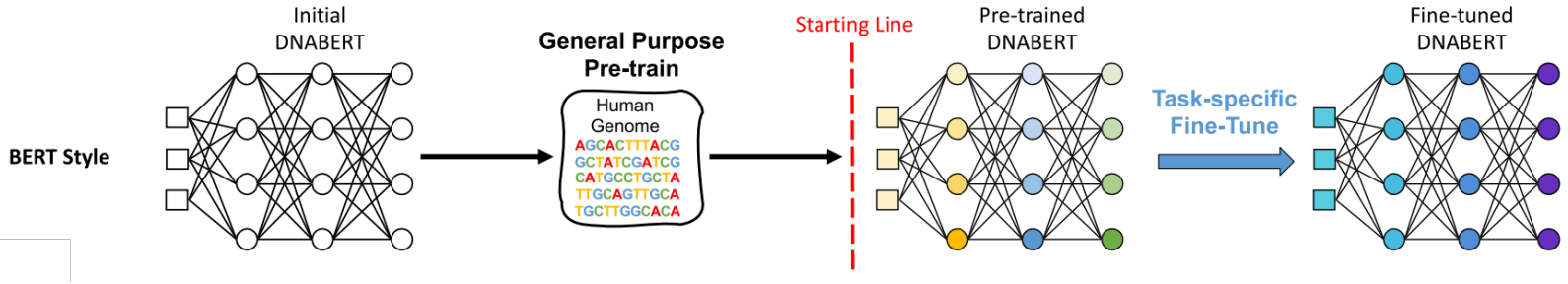
DNA/Genomic Language Modeling

The large scale of microbial contigs or whole genomes, ranging from 0.5 to 10 million base pairs, often exceeds transformer context windows, necessitating the use of additional techniques like CNNs to analyze genes.

Tasks:
predict gene function (Classification and Prediction)
predict proximal and core promoter regions (Classification and Prediction)
identify transcription factor binding regions (Understanding Tasks)
figure out important regions and sequence motifs (Information Extraction)
Generate sequences with potential specific functions (Generation Tasks)

Datasets:
PubMLST
Microbiome Database (MDB)
CNGBdb
Environmental Microbiome Database (ES-DB)

DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome

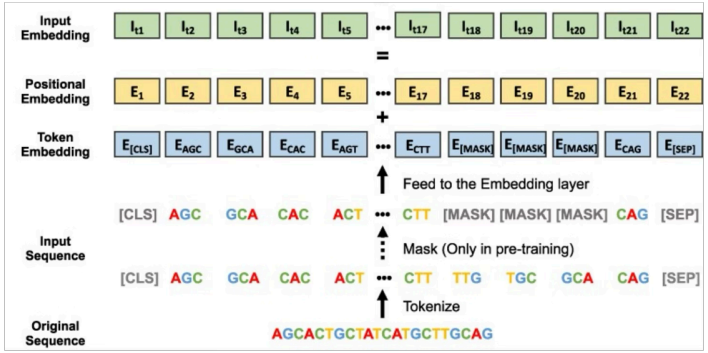


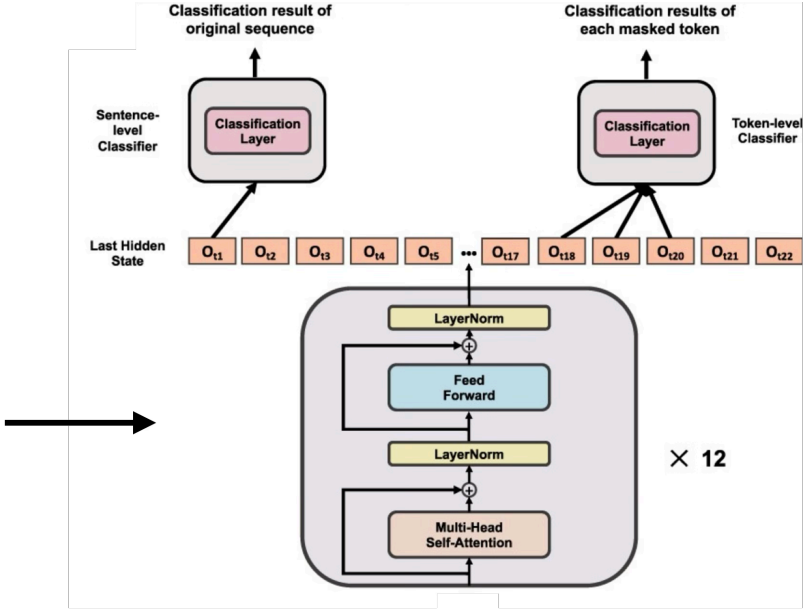
DNABERT adopts general-purpose pre-training which can then be fine-tuned for multiple purposes using various task-specific data.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, Michal Linial, ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics, Volume 38, Issue 8, March 2022

DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome

DNABERT addresses the technical challenges of traditional CNN and RNN models, which struggle to effectively capture global context and long-range dependencies in long sequences.

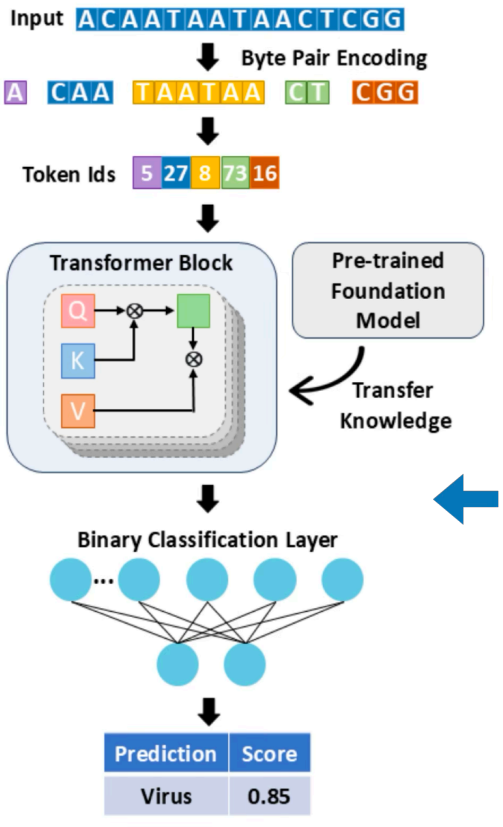




Compared to traditional methods, DNABERT achieves superior performance and better cross-organism adaptability through its pre-training and fine-tuning approach, even in data-scarce scenarios.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, Michal Linial, ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics, Volume 38, Issue 8, March 2022

ViraLM: Empowering Virus Discovery through the Genome Foundation Model



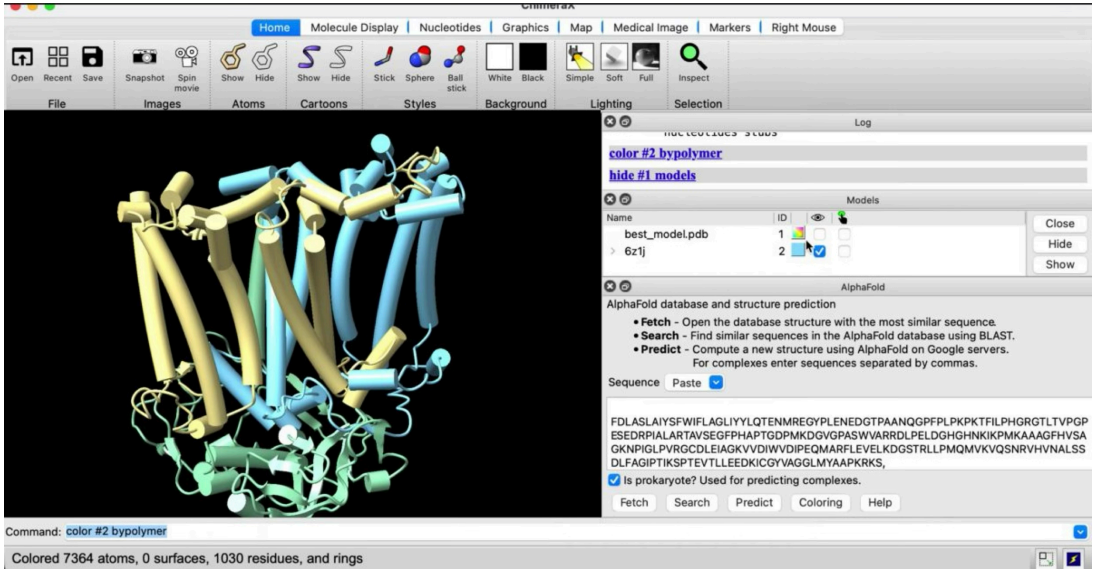
Detecting viruses in metagenomic data :
Limited Reference Sequences
Difficulty in Identifying Short Viral Fragments
Inconsistency in Search Results
Limitations of Protein-Based Methods

The main architecture of ViraLM:
a pre-trained Transformer block from DNABERT-2
a fine-tuned binary classifier layer for virus classification.

Future Directions

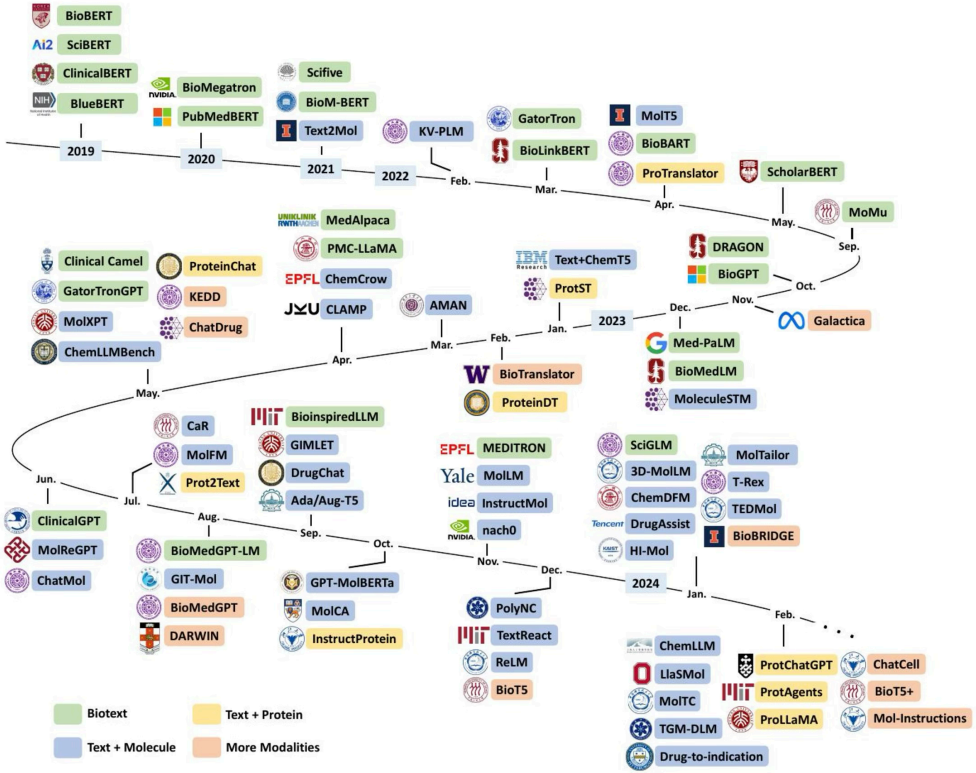
Low-resource learning: lack of annotated data.

AI for science: user-friendly assistant tools with lower barriers to entry; unleash human researcher productivity.



AlphaFold – run from ChimeraX

Cross-modal processing: bridging biomolecules and natural language or different forms



Leveraging Biomolecule and Natural Language through Multi-Modal Learning: A Survey

Takeaways

1. Protein and DNA sequences resemble natural language, enabling NLP techniques to analyze complex dependencies in microbiomes and metagenomic data.
2. ProteinBERT leverages transformer architecture to predict protein functions and resistance mechanisms, outperforming traditional methods and demonstrating interpretability in bioinformatics.
3. The large scale of microbial genomes necessitates advanced techniques like CNNs and DNABERT, which effectively capture long-range dependencies and provide contextualized representations of DNA sequences.
4. ViraLM builds on DNABERT-2 to enhance virus classification by recognizing subtle differences in nucleotide sequences, demonstrating the effectiveness of transfer learning.

Thank you！

Reference

1. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics. 2021 Aug 9;37(15):2112-2120. doi: 10.1093/bioinformatics/btab083. PMID: 33538820; PMCID: PMC11025658.

2. Pei, Qizhi, Wu. Leveraging Biomolecule and Natural Language through Multi-Modal Learning: A Survey. arXiv preprint arXiv:2403.01528

3. Peng, Cheng et al. “ViraLM: Empowering Virus Discovery through the Genome Foundation Model.” Bioinformatics (Oxford, England), btae704. 23 Nov. 2024, doi:10.1093/bioinformatics/btae704

4. Y. Pang and B. Liu, "SelfAT-Fold: Protein Fold Recognition Based on Residue-Based and Motif-Based Self-Attention Networks," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 3, pp. 1861-1869, 1 May-June 2022, doi: 10.1109/TCBB.2020.3031888.

5. Transforming the Language of Life: Transformer Neural Networks for Protein Prediction Tasks
Ananthan Nambiar, Simon Liu, Mark Hopkins, Maeve Heflin, Sergei Maslov, Anna Ritz
bioRxiv 2020.06.15.153643; doi: https://doi.org/10.1101/2020.06.15.153643

6. Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, Michal Linial, ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics, Volume 38, Issue 8, March 2022.

7. Kanami Yagimoto, Shion Hosoda, Miwa Sato, Michiaki Hamada, Prediction of antibiotic resistance mechanisms using a protein language model, *Bioinformatics*, Volume 40, Issue 10, October 2024

Reference

1. Ma, Y., Guo, Z., Xia, B. et al. Identification of antimicrobial peptides from the human gut microbiome using deep learning. Nat Biotechnol 40, 921–931 (2022).

2. Cai Chen, Shu-Le Li, Yao-Yang Xu, Jue Liu, David W. Graham, Yong-Guan Zhu, Characterising global antimicrobial resistance research explains why One Health solutions are slow in development: An application of AI-based gap analysis, Environment International, Volume 187, 2024,

3. Mingxin Hou, Xiaowen Zhong, Ouyang Zheng, Qinxiu Sun, Shucheng Liu & Mingxin Liu. (2025) Innovations in seafood freshness quality: Non-destructive detection of freshness in Litopenaeus vannamei using the YOLO-shrimp model. Food Chemistry 463, pages 141192.

4. Mohammad H. Alshayeji, Silpa ChandraBhasi Sindhu, Sa'ed Abed, Viral genome prediction from raw human DNA sequence samples by combining natural language processing and machine learning techniques, Expert Systems with Applications, Volume 218, 2023

5. Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

6. Danielle Miller, Ofir Arias, David Burstein, GeNLP: a web tool for NLP-based exploration and prediction of microbial gene function, Bioinformatics, Volume 40, Issue 2, February 2024